

Sequence Heterogeneity Accelerates Protein Search for Targets on DNA

Alexey A. Shvets and Anatoly B. Kolomeisky*

Rice University, Department of Chemistry and Center for Theoretical Biological Physics, Houston, Texas 77005, USA

(Dated: July 23, 2015)

The process of protein search for specific binding sites on DNA is fundamentally important since it marks the beginning of all major biological processes. We present a theoretical investigation that probes the role of DNA sequence symmetry, heterogeneity and chemical composition in the protein search dynamics. Using a discrete-state stochastic approach with a first-passage events analysis, which takes into account the most relevant physical-chemical processes, a full analytical description of the search dynamics is obtained. It is found that, contrary to existing views, the protein search is generally faster on DNA with more heterogeneous sequences. In addition, the search dynamics might be affected by the chemical composition near the target site. The physical origins of these phenomena are discussed. Our results suggest that biological processes might be effectively regulated by modifying chemical composition, symmetry and heterogeneity of a genome.

Usage: Secondary publications and information retrieval purposes.

PACS numbers: May be entered using the `\pacs{#1}` command.

Structure: You may use the `\description` environment to structure your abstract; use the optional argument of the `\item` command to give the category of each item.

Many biological processes are initiated by proteins binding the specific target sequences on DNA. In particular, this process is responsible for transferring and maintaining the genetic information contained in DNA [1–3]. It was recognized long time ago that finding these specific binding sites could be quite a complicated task due to large number of other nonspecific sites ($\simeq 10^6 - 10^9$) and low concentration of relevant proteins. But experiments suggest that many proteins find their targets much faster than expected from 3D bulk diffusion estimates [4–8]. This surprising phenomenon is known as a *facilitated diffusion*. A significant progress in explaining facilitated diffusion processes has been achieved in recent years due to multiple experimental and theoretical advances [5–32]. However, the detailed mechanisms of the protein search for targets on DNA remain not well understood [7, 8, 23].

It is now widely accepted that proteins searching for the specific binding sites on DNA at some conditions might alternate between 3D and 1D search modes [5, 7–9, 11]. This means that the protein molecule binds non-specifically to DNA, then slides along the chain, unbinds and repeats the scanning cycle several times until it finds the target. Recent single-molecule experiments that can visualize the dynamics of individual molecules support this picture [12, 15, 16, 21, 27, 28]. These observations also underline the critical role of protein-DNA interactions in the facilitated diffusion. Since DNA molecule is a heterogeneous biopolymer, the sequence symmetry and its chemical composition must be an important factor in the protein search for targets. However, how specifically the sequence heterogeneity influences the protein search dynamics remains a controversial problem.

The protein search on the random DNA sequences have been theoretically investigated before [7, 33]. Comparing

this process with a motion in the random potential, it was shown that the heterogeneous character of the chain leads to the larger search times in comparison with a homogeneous case. But later it was argued that this result is not applicable to the protein search [23]. It is just an artifact of the continuum approximation, which assumed that the protein can reach the target only via DNA sliding, neglecting 3D associations and dissociations events [23]. A more advanced computational study of the sequence heterogeneity also found that it usually slows down the facilitated diffusion by creating traps [34]. But it was also suggested that the properly positioned traps in the funnel shape near the target can accelerate the protein search [34]. At the same time, it is not clear if such funnel distributions are observed in real systems. Furthermore, recent theoretical studies of Lukatsky and coworkers [35–38] suggested that the sequence symmetry creates additional effective interactions between DNA and protein molecules. Using methods of equilibrium statistical mechanics, it was found that more homogeneous segments of DNA effectively attract proteins stronger than the heterogeneous segments. However, the role of these effective interactions in the protein search for targets on DNA has not been tested yet.

In this article, we present a theoretical approach that allows us to investigate explicitly the effect of sequence heterogeneity in the protein search for targets on DNA. It is based on a discrete-state stochastic method which takes into account the most relevant physical-chemical processes of the protein search by analyzing first-passage events in the system [23, 30]. The advantage of this method is that it provides a full analytical description of the facilitated diffusion. One of the main results of this approach is a development of the general dynamic phase diagram for the target search [23]. Three dynamic search regimes were identified depending on the different length scales in the system. For protein slid-

* tolya@rice.edu

ing length λ larger than the size of the DNA chain L , the protein molecule always stays on DNA and performs 1D search with a random-walk dynamics. This leads to the quadratic scaling of the search times as a function of the DNA length. When the sliding length is smaller than the length of DNA but larger than the target size ($1 < \lambda < L$), the protein is searching by combining 3D and 1D motions. In this sliding regime, the linear scaling of the search times is observed. A different dynamic phase is found for the case of the sliding length smaller than the target size, $\lambda < 1$. Here the search is accomplished only via 3D associations and dissociations events without sliding along the DNA molecule. This also leads to the linear scaling in the search times as a function of the DNA length.

In our model, we consider a single DNA molecule with $L+1$ binding sites and a single protein molecule, as shown in Fig. 1. One of the binding sites is a target, and for convenience we put it in the middle of the chain, i.e., $m = L/2 + 1$. To model the sequence heterogeneity, we assume that each monomer in the DNA chain can be in one of two chemical states, A or B (see Fig. 1). When the protein is bound to the segment A (B) it interacts with energy ε^A (ε^B), and $\varepsilon = \varepsilon^A - \varepsilon^B \geq 0$. This means that the protein attracts stronger to the B sites. The protein molecule can diffuse along DNA with the rate $u_A \equiv u$ ($u_B = ue^{-\varepsilon}$, where ε is measured in $k_B T$ units). Here we assume that, independently of the chemical state of their neighbors, moving out of the sites A are characterized by the rate u_A , while the diffusion out of the sites B is given by u_B . The protein search starts in the solution that we label as a state 0. Then the protein molecule can bind to any site A or B on DNA with the corresponding rates $k_{on}^A \equiv k_{on}$ or $k_{on}^B = k_{on}e^{\theta\varepsilon}$. Similarly, the dissociations from the DNA chain are described by the rates $k_{off}^A \equiv k_{off}$ and $k_{off}^B = k_{off}e^{(\theta-1)\varepsilon}$. Here the parameter $0 \leq \theta \leq 1$ specifies how the protein-DNA interaction energy is distributed between the association and dissociation transitions. We also assume that the binding to the target is given by $k_{on}^T = k_{on}$. To test the effect of the sequence symmetry and heterogeneity we consider the protein search on two different types of the DNA molecules: see Figs. 1b and 1c. One of them consists of two homogeneous segments of only A and only B subunits separated by the target (Fig. 1b). Another one is the biopolymer with alternating A and B sites, as presented in Fig. 1c. The block copolymer (Fig. 1b) has a more homogeneous sequence, while the alternating polymers (Fig. 1c) are more heterogeneous. It is important to note that in both cases the overall interaction between the protein and DNA is the same (the overall chemical composition in both cases is identical), and thus our analysis probes *only* the effect of the heterogeneity. This is different from previous computational studies [34].

To describe the target search dynamics, let us introduce a function $F_n(t)$, which is defined as a first-passage probability to reach the target, if at $t = 0$ the protein was at the site n ($n = 1, 2, \dots, L+1$ corresponds to the

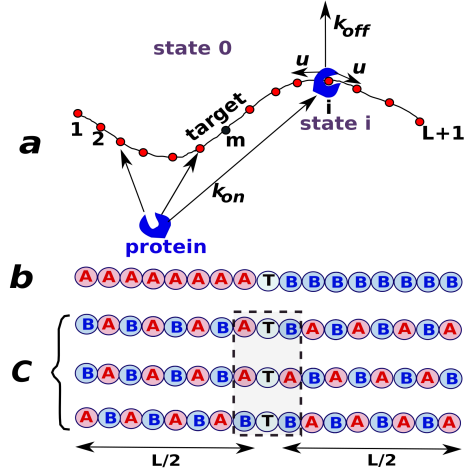


FIG. 1. a) A general scheme of the protein search. The DNA chain consists of L nonspecific binding sites and one specific site that is a target for the search. A protein, coming from the solution, can bind to any site on DNA with the association rate per one segment given by $k_{on}^{(i)}$ with $i = A$ or B . When attached, the protein can diffuse along the DNA with the rate u_i ($i = A$ or B), and it can dissociate into the solution with the rate $k_{off}^{(i)}$ ($i = A$ or B). The search is finished when the protein binds to the target site at the position $m = L/2 + 1$. b) A fully symmetric AB block copolymer DNA sequence. c) Pseudo-random alternating sequences with different compositions near the target.

starting DNA and $n = 0$ is for the bulk solution). The temporal evolution of this quantity can be described by the backward master equations [23],

$$\frac{dF_n(t)}{dt} = u_n[F_{n-1} + F_{n+1}] + k_{off}^{(n)}F_0(t) - (2u_n + k_{off}^{(n)})F_n(t), \quad (1)$$

for $1 \leq n \leq L+1$, while in the solution we have

$$\frac{dF_0(t)}{dt} = \sum_{n=1}^{L+1} k_{on}^{(n)}F_n(t) - F_0(t) \sum_{n=1}^{L+1} k_{on}^{(n)}. \quad (2)$$

It is convenient to analyze these equations in the Laplace space using a transformation $\tilde{F}_n(s) = \int_0^\infty F_n(t) e^{-st} dt$. Then all probabilities can be found explicitly, which leads to the full dynamic description of the search process. The details of the calculations are presented in the Supplementary Material. More specifically, the mean first-passage time to reach the target starting from the solution is given by $T_0 \equiv -\frac{\partial \tilde{F}_0(s)}{\partial s}|_{s=0}$, and other dynamic properties can be also written explicitly. This framework allows us to compare the search dynamics on DNA with different sequences.

In the case of more homogeneous block copolymer sequence (see Fig. 1b), the mean search times are equal to

$$T_0 = \frac{k_{off} + k_{on}[(L/2 - P^A) + e^\varepsilon(L/2 - P^B)]}{k_{on}k_{off}(1 + P^A + e^{\theta\varepsilon}P^B)}, \quad (3)$$

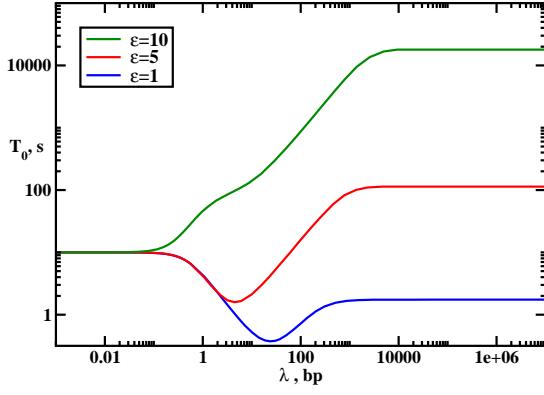


FIG. 2. Average times to find the target for block copolymer DNA sequence as a function of the scanning length $\lambda = \sqrt{u/k_{off}}$. The transition rates are: $u = 10^5 \text{ s}^{-1}$ and $k_{on} = 0.1 \text{ s}^{-1}$. The DNA length is $L = 1000$, and we vary the energy difference ε (in units of $k_B T$) for the interaction between the protein and A and B subunits on DNA.

where

$$P^{(i)} = \frac{x_i^{1-L/2} - x_i^{1+L/2}}{(1 - x_i)(x_i^{1+L/2} + x_i^{-L/2})}, \quad (4)$$

$$x_i = \frac{2u_i + k_{off}^{(i)} - \sqrt{(2u_i + k_{off}^{(i)})^2 - 4u_i^2}}{2u_i}, \quad (5)$$

for $i = A$ and B . The results are presented in Fig. 2. Again, three dynamic search phases are clearly observed. Increasing the strength of interactions with B subunits make the search in the random-walk regime much slower. This is because the protein gets effectively trapped on B sites for $\lambda > L$.

Similar expressions for the mean first-passage times can be found for AB alternating DNA chains, as shown in the Supplementary Material. Here we use the pseudo-random alternating sequences, mimicking the real random situations, because the analytical results can be obtained for them. But we tested this approximation in computer Monte Carlo simulations by generating random sequences, and one can see from Fig. 3 that this assumption is fully justified. Another interesting observation from Fig. 3 is that the chemical composition near the target might also affect the search dynamics. This can be found only for the intermediate sliding regime ($1 < \lambda < L$) because in this case the probability fluxes to the target site from the solution and from the DNA are comparable. Modifying the composition of the sites near the target can change the amount of the flux coming from the DNA chain. The flux is larger for BTB sequences (2 B subunits around the target), leading to the smaller search times. This is because the protein molecule attracts stronger to B sites and it has a higher probability to be found here and eventually to go the target. At the same time the flux is smaller for ATA sequences (2 A

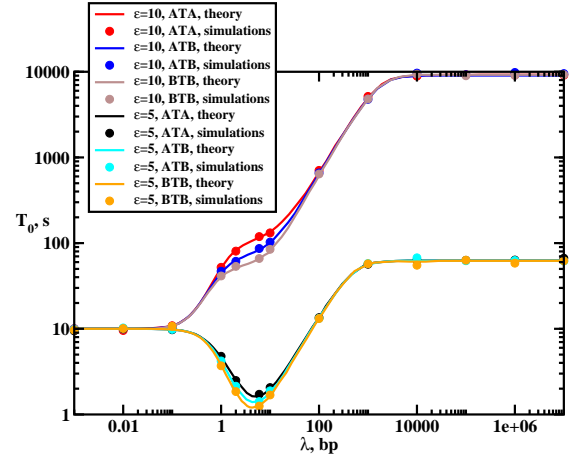


FIG. 3. Comparison of the search times for alternating sequences with random sequences generated in Monte Carlo computer simulations. The transition rates are: $u = 10^5 \text{ s}^{-1}$ and $k_{on} = 0.1 \text{ s}^{-1}$. The DNA length is $L = 1000$, the loading parameter is $\theta = 0.5$, and two different interaction strengths, $\varepsilon = 10$ and $\varepsilon = 5$, are probed.

subunits around the target) with weaker interactions to A sites, which yields slower search dynamics. For ATB sequences, as expected, the intermediate dynamics is observed.

Now we can quantify the effect of sequence heterogeneity in the protein search for the specific binding sites on DNA. The results in Fig. 4 present a ratio of the search times for block copolymer sequences, which are less heterogeneous, and for various alternating sequences, which are more heterogeneous, as a function of the sliding length on DNA. One can see that the effect of the sequence heterogeneity depends on the nature of the dynamic search phase. In the jumping regime ($\lambda < 1$), the symmetry of the sequence does not play any role. This is because in this case the process is taking place only via associations and dissociations (3D search), and the structure of the DNA chain is not important. The situation is different for the intermediate sliding regime (3D+1D search, $1 < \lambda < L$) where in most cases the search on alternating sequences is faster. This can be explained by noting that the search time in this dynamic phase is proportional to L/λ [23], which gives the average number of cycles before the protein can find the target. In the block copolymer sequence the protein mostly comes to the target from the B segment because of stronger interactions. In the alternating sequences the protein can reach the target from both sides. It can be shown analytically (see the Supplementary Material) that the scanning length on the alternating segment is larger than the scanning length for the B segment, i.e., $\lambda_{AB} > \lambda_B$. Then the search time is obviously faster for the alternating sequence because $L/\lambda_{AB} < L/\lambda_B$. The only deviation from this picture is found in ATA sequences where for small range of parameters the search is slower than in the block copolymer sequence. The effect of the chemical composition near

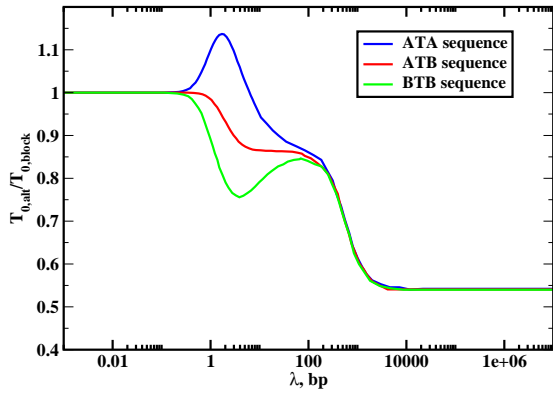


FIG. 4. The ratio of the search times for the alternating DNA sequences and for the block copolymer DNA sequences as a function of the scanning length $\lambda = \sqrt{u/k_{off}}$. Three different chemical compositions near the target are distinguished, namely, *ATA*, *ATB*, *BTB*. The transition rates are: $u = 10^5 \text{ s}^{-1}$ and $k_{on} = 0.1 \text{ s}^{-1}$. The DNA length is $L = 1000$, the loading parameter is $\theta = 0.5$, and the energy difference of interactions for the protein with *A* and *B* sites is $\varepsilon = 5$.

the target, as discussed above, is responsible for this. In the random-walk regime (1D search, $\lambda > L$), the effect of the sequence heterogeneity is even stronger: the protein molecule finds the specific binding site up to 2 times faster for more heterogeneous DNA chains. To understand this behavior, we note that in this case the mean first-passage time to reach the target is a sum of residence times on the DNA sites. Because the target is in the middle of the chain, the mean time to reach the target from the block copolymer sequence will be $T_0 \simeq (L/4)\tau_B$, where τ_B is the residence time at the site *B*. The average starting position of the protein is $L/4$ sites away from the target. For the alternating sequences, the average distance to the target is the same, but the chemical composition of intermediate sites is different, yielding, $T_0 \simeq (L/8)\tau_B + (L/8)\tau_A$. Obviously, the protein spends much less time on *A* subunits, and this leads to faster

search for alternating DNA sequence. For $\tau_A \ll \tau_B$ this also explains the factor of 2 in the search speed. In this case, the *B* subunits can be viewed as traps. Thus, in dynamic phases where the structure of DNA is important the sequence heterogeneity almost always accelerates the protein search for targets.

In conclusion, we presented a theoretical analysis of DNA sequence symmetry and heterogeneity in the protein search process. Using analytical solutions of the discrete-state stochastic approach that accounts for most important physical-chemical processes in the system, we obtained a full description of the search dynamics. It is found that the sequence heterogeneity is a crucial factor in the facilitated diffusion. Unlike the previous theoretical and computational models, our approach predicts that the sequence heterogeneity mostly accelerates the search. The mechanisms of this phenomenon depend on the nature of the search regime. It is either the smaller number of search cycles or the smaller number of trapping sites on the path to the target. We also found that in the dynamic phase where the specific binding site can be reached from the solution and from the DNA chain, the chemical composition near the target might influence the search dynamics. The search is faster if the target is surrounded by the subunits which interact stronger with the protein, providing it more opportunities to reach the target. Our theoretical results not only clarify the fundamental physics of the protein search dynamics, but they also suggest that the biological processes can be effectively regulated by modifying the sequence symmetry and heterogeneity in DNA, as well as the chemical composition near the targets. Experiments to test these predictions should provide a better understanding of the microscopic mechanisms of complex biological processes.

The work was supported by the Welch Foundation (Grant C-1559), by the NSF (Grant CHE-1360979), and by the Center for Theoretical Biological Physics sponsored by the NSF (Grant PHY-1427654).

-
- [1] Alberts, B., et al. *Molecular Biology of Cell* 6th ed. (Garland Science, New York, 2014).
 - [2] H. Lodish, et al. *Molecular Cell Biology* 6th ed. (W.H. Freeman, New York, 2007).
 - [3] R. Phillips, J. Kondev, and J. Theriot, *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2012).
 - [4] A.D. Riggs, S. Bourgeois, M. Cohn, *J. Mol. Biol.* **48**, 67 (1970).
 - [5] S. E. Halford and J.F. Marko, *Nucl. Acid Res.* **32**, 3040 (2004).
 - [6] B. van den Broek, M.A. Lomholt, S.-M. Kalisch, R. Metzler and G.J.L. Wuite, *Proc. Natl. Acad. Sci. USA* **105**, 15738 (2008).
 - [7] L.A. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J.S. Leith, and A. Kosmrlj, *J. Phys. A: Math. Theor.* **42**, 434013 (2009).
 - [8] A.B. Kolomeisky, *Phys. Chem. Chem. Phys.* **13** 2088 (2011).
 - [9] O.G. Berg, R.B. Winter, and P.H. von Hippel, *Biochemistry* **20**, 20, 6948 (1981).
 - [10] O.G. Berg and P.H. von Hippel, *Ann. Rev. Biophys. Biophys. Chem.* **14**, 131 (1985).
 - [11] R.B. Winter, O.G. Berg, and P.H. von Hippel, *Biochemistry* **20**, 6961 (1981).
 - [12] D.M. Gowers, G.G. Wilson, G.G. and S.E. Halford, *Proc. Natl. Acad. Sci. USA* **102**, 15883 (2005).
 - [13] J. Iwahara, M. Zweckstetter, and G.M. Clore, *Proc. Natl. Acad. Sci. USA* **103**, 15062 (2006).
 - [14] G. Kolesov, Z. Wunderlich, O.N. Laikova, M.S. Gelfand, and L.A. Mirny, *Proc. Natl. Acad. Sci. USA* **104**, 13948 (2007).
 - [15] Y.M. Wang, R.H. Austin, and E.C. Cox, *Phys. Rev. Lett.*

- 97**, 048302 (2006).
- [16] J. Elf, G.-W Li, and X.S. Xie, *Science* **316**, 1191 (2007).
 - [17] A. Tafvizi, F. Huang, J.S. Leith, A.R. Fersht, and L.A. Mirny, *Biophys. J.* **95**, L1-L3 (2008).
 - [18] T. Hu, A. Yu. Grosberg and B. I. Shklovskii, *Biophys. J.* **90**, 2731 (2006)
 - [19] D.C. Rau and N.Y. Sidorova, *J. Mol. Biol.* **395**, 408 (2010).
 - [20] D.R. Larson, D. Zenklusen, B. Wu, J.A. Chao, and R.H. Singer, *Science* **332**, 475 (2011).
 - [21] P. Hammar, P. Leroy, A. Mahmutovic, E.G. Marklund, O.G. Berg, and J. Elf, *Science* **336**, 1595 (2012).
 - [22] L. Zandarashvili, D. Vuzman, A. Esadze, Y. Takayama, D. Sahu, Y. Levy, and J. Iwahara, *Proc. Natl. Acad. Sci. USA* **109**, E1724 (2012).
 - [23] A. Veksler and A.B. Kolomeisky, *J. Phys. Chem. B* **117**, 12695 (2013).
 - [24] A. Marcovitz and Y. Levy, *Biophys. J.* **104**, 2042 (2013).
 - [25] E.F. Koslover, M.A.D. de la Rosa, and A.J. Spakowitz, *Biophys. J.* **101**, 856 (2011).
 - [26] M. Sheinman, O. Benichou, Y. Kafri, and R. Voituriez, *Rep. Prog. Phys.* **75**, 026601 (2012).
 - [27] M.P. Landry, X. Zou, L. Wang, W.M. Huang, K. Schul-
ten, and Y.R. Chemla, *Nucl. Acids Res.* **41**, 2416 (2013).
 - [28] A. Tafvizi, F. Huang, A.R. Fersht, L.A. Mirny, and A.M. van Oijen, *Proc. Natl. Acad. Sci. USA* **108**, 563 (2011).
 - [29] J.S. Leith, A. Tafvizi, F. Huang, W.E. Uspal, P.S. Doyle, A.R. Fersht, L.A. Mirny, and A.M. van Oijen, *Proc. Natl. Acad. Sci. USA* **109**, 16552 (2012).
 - [30] A.B. Kolomeisky and A. Veksler, *J.Chem.Phys.* **136**, 125101 (2012) .
 - [31] A. Esadze, C.A. Kemme, A.B. Kolomeisky, and J. Iwahara, *Nucl. Acids Res.* **42** 7039 (2014).
 - [32] M. Bauer and R. Metzler, *PLOS One* **8** e53956 (2013).
 - [33] T. Hu and B. I. Shklovskii, *Phys. Rev. E* **74**, 021903 (2006).
 - [34] C. A. Brackley, M. E. Cates, and D. Marenduzzo, *Phys. Rev. Lett.* **109**, 168103 (2012).
 - [35] A. Afek, I. Sela, N. Musa-Lempel, and D.B. Lukatsky, *Biophys. J.* **101** 2465 (2011).
 - [36] A. Afek and D.B. Lukatsky, *Biophys. J.* **102** 1881 (2012).
 - [37] A. Afek and D.B. Lukatsky, *Biophys. J.* **105** 1653 (2013).
 - [38] A. Afek, J. L. Schipper, J. Horton, R. Gordn, and D. B. Lukatsky, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17140 (2014)